DOCUMENT RESUME

ED 045 709 TM 000 278

AUTHOR Murray, James R.; Wiley, David E.

TITLE New Statistical Techniques for Evaluating Longitudinal Models.

INSTITUTION Chicago Univ., Ill.
SPONS AGENCY Early Education Research Center, Chicago, Ill.
PUB DATE Sep 70
NOTE 13p.; Paper presented as part of the Symposium "Models and Methods for the Study of the Life Cycle", given at the American Psychological Association Convention, Miami Beach, Florida, September 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.75
DESCRIPTORS Behavioral Sciences, *Data Analysis, *Evaluation Techniques, Feedback, Goodness of Fit, *Longitudinal Studies, *Mathematical Models, Probability, Program Evaluation, *Research Methodology, *Statistical Data

ABSTRACT

A basic methodological approach in developmental studies is the collection of longitudinal data. Behavioral data can take at least two forms, qualitative (or discrete) and quantitative. Both types are fallible. Measurement errors can occur in quantitative data and measures of these are based on error variance. Qualitative or discrete data can contain misclassification errors, and these are expressed as probabilities of misclassification. Statistical models for psychological data must take these differences into account. A simple sequence is presented as an example of a qualitative model, while disengagement is the model given as an example for quantitative data. These examples, which are special cases of more general problems, lead to an outline of the general nature of the qualitative and quantitative models. The primary concern here is to develop statistical models which permit the investigation of structure in fallible longitudinal data. Statistical descriptions of the simple sequence and disengagement models are included in the appendix. (CK)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED045709

NEW STATISTICAL TECHNIQUES FOR EVALUATING LONGITUDINAL MODELS

James R. Murray and David E. Wiley
University of Chicago

Presented as part of the Symposium
"Models and Methods for the Study of the Life Cycle"
Chairman, Bernice L. Neugarten

given at the

American Psychological Association Convention
Miami Beach, Florida, September, 1970

This research is supported in part by Early Education Research Center (pursuant to a contract with the Office of Education, U. S. Department of Health, HEW) and the current NSF Grant, GS - 1930.

000 278
M

Part I: General Description

A basic methodological approach in developmental studies is the collection of longitudinal data, i.e. observations on the same Ss at multiple points in time. Two often asked questions of such data are:

1) Are there invariant sequences of behavioral phenomenon?
2) What are the processes which cause variables to change over time? Or, what controls age related changes in variables?

There are obviously a number of additional questions which can be asked of longitudinal data. However, these particular problems are quite general. These questions can be found in studies of childhood as well as in studies of old age. Furthermore, such questions can provide the statistician working with developmental data with a useful starting point. We have approached these questions as problems in statistical model building.

Behavioral data can be of (at least) two forms, the first, qualitative or discrete, and the second, quantitative. Qualitative or discrete data is generally characterized by observations which are categorized into one of a number of mutually exclusive classes. Variables such as occupation, cognitive stage, or marital status are discrete or qualitative. Quantitative data, on the other hand, arise from observations which yield measures on a ratio scale. Examples of variables which are quantitative are weight, height and true-scale scores.

Each of these two kinds of data is fallible, however. Quantitative data can have measurement error in it, and classical test reliability

(J. Ward Keesling, Department of Education, UCLA, has been responsible for much of the work on quantitative models.)

theory has been developed to meet this statistical problem. The measurement error in quantitative data is itself quantitative. Because of this, a measure of the noise or error in a quantitative variable is based on the error variance. Qualitative data is also fallible, and this is referred to as misclassification error. Piaget's framework can provide an example of a misclassification error. There can be a nonzero probability that a child could really be at the preoperational stage of cognitive development but be observed or classified as being at the level of concrete operations. Misclassification error for discrete data is expressed as probabilities of misclassification, as opposed to error variance which characterizes quantitative data.

Statistical models for psychological data should take into account the difference between quantitative and qualitative or discrete data. Furthermore, each kind of model includes measurement error parameters appropriate to the data form. The inclusion of such error parameters is significant because it renders inaccurate some statistical estimation procedures. For example, the simple least square procedure used in regression analysis will not yield correct regression weights when quantitative independent variables are measured with error. Models have been developed for longitudinal data which contains measurement error.

A Simple Sequence: An example of a qualitative model. Cognitive development in the child has been seen, especially by Piaget, as essentially a sequential process of passing through various qualitatively distinct stages of cognitive organization. The verification of such an assertion requires longitudinal data. Such data must meet the constraint that each child pass through or possess cognitive stages in the proper order. An observed sequence of stages for a sample of children, however, will probably show

some children who do not change in the predicted way. These observed patterns which are theoretically inadmissible could be due simply to misclassification error. We have developed a model for this problem which includes misclassification error for each stage or qualitative category. This model will allow the observed probabilities to be 'incorrect' yet have the underlying or latent process follow a strict sequence at the same time. To obtain estimates of both the misclassification errors on the one hand, and the underlying transition rates between stages on the other hand, the method of maximum likelihood is used. Furthermore, the overall goodness of fit of the sequence-plus-misclassification error model is tested.

Disengagement: An example for quantitative data. The problem of disentangling antecedent-consequent relationships among variables which cannot be experimentally controlled is one which students of the life-cycle regularly face. The 'disengagement hypothesis' of Cumming & Henry is centrally concerned with such a problem. The question behind this particular hypothesis concerns the relationship between the psychological and social involvement of individuals as they enter the period of retirement and old age. The 'bare-bones' of this hypothesis is that aging naturally entails a withdrawal from society which is preceded or anticipated, on the individual level, by an increasing psychological focus on the self. Verification of such a hypothesis can be accomplished by collection of longitudinal data regarding the degree of social interaction as well as data regarding the degree of ego involvement in the world of people and objects outside of the self. Once such data are available, problems of analysis come to the foreground. There are two major issues here:

1) The data will have measurement error. That is to say, the measures of social involvement and psychological involvement will each be subject to error. In this case, classical linear regression does not yield correct estimates of the true or latent relationship between the variables.

2) The underlying relationship between psychological involvement and social involvement must be directly expressed in a structural model. The basic antecedent-consequent relation can be expressed by letting the level of social involvement at a given time be linearly dependent on the level of psychological involvement at the immediately preceding time. This linear dependence over time is postulated to hold on the latent or true part of the variables. That is, measured social involvement is not a simple function of measured psychological involvement since each measure has error.

We have assumed that these two variables can be quantitatively measured. The model which includes structural equations and measurement error can be estimated by using the method of maximum likelihood. A test of goodness of fit is given by a likelihood ratio.

General Comments: The example of a qualitative model, the simple sequence, and the example of a quantitative model, the disengagement hypothesis, are only special cases of more general problems. Our work is primarily concerned with developing statistical models which enable one to investigate structure in fallible longitudinal data. Since many substantive problems require different forms of data, for example, bone growth vs. ego development, we have tried to extend structural models to

qualitative data as well as quantitative data. The general nature of the models we are considering can be described as follows.

The general qualitative model is conceptually related to Lazarsfeld's latent structure analysis. There are two major differences:

1) Items, the qualitative variables, have as many misclassification parameters (probabilities) as they have categories of response in our models. For example, dichotomous items, e.g. yes/no, have two parameters--one for each corresponding latent state. In Lazarsfeld's system, on the other hand, items often are given only one misclassification parameter.
2) The latent classes of Lazarsfeld, which each have a latent probability, are highly restricted in our own models. A given latent class probability in Lazarsfeld's model is expressed as a function of various latent probability parameters in our models. Our approach to parameterization is necessary in testing hypotheses which involve structured processes among latent classes.

The general quantitative model is regarded as a covariance structure model. The quantitative measures are assumed to be distributed as multivariate normal vectors and our application of maximum likelihood solutions corresponds to what the econometricians call a full-information method for solving structural regression problems. The quantitative model allows longitudinal data to be structured a number of different ways. For example:

1) Feedback - 2 or more variables which each determine one another over time can be examined;

2) Multiple cause-effect - the nature of simple interdependence of many variables over time can be studied; and

3) Systems of processes - the nature of complex chains of dependencies among variables can be studied with longitudinal data.

In conclusion, some comment on the general substantive relevance of these kinds of statistical models should be made. Many developmental psychologists claim that few, if any, tenable theories have been available in this general field. Thus, most data analysis is really a hunting expedition rather than a process of rigorous confirmatory study. By the looks of things, this statement is on the whole descriptively accurate. The role of our statistical models is not that of the divining rod for scientific discoveries. It is our belief, however, that successful scientific hunting is found to be so in confirmatory analysis. Our models are built to do confirmatory analysis in a way not previously available.

Part II: Statistical Description

The Qualitative Model: The general form of the models for qualitative data is given by (1).

(1) $$\underline{p} = Q\underline{\Pi}\ ,$$ where

$\underline{p}$ is an $m \times 1$ vector of observed or manifest probabilities.

Q is a $m \times m$ matrix of misclassification probabilities.

$\underline{\Pi}$ is an $m \times 1$ vector of latent class probabilities.

Q has the following characteristics:

A) For each of the k separate items used at a given time, there is a separate matrix of misclassification probabilities, say Q_i, $i = 1,\ldots,k$.

B) Each Q_i is a mtrix of conditional probabilities of the form

(2) $$q_i^{(j\ell)} = p(r_j|\rho_\ell),$$

where r_j is the j^{th} manifest or observed response to item i, $j = 1,\ldots,J_i$

ρ_ℓ is the ℓ^{th} true or latent category for item i, $\ell = 1,\ldots, L_i$, $(L_i = J_i)$; and

(3) $$\sum_{j=1}^{j} q_i^{(j\ell)} = 1.$$

C) Conditional independence of the response errors is assumed so that, at the t^{th} time of measurement;

(4) $$Q_t = Q_1 \otimes Q_2 \otimes \cdots \otimes Q_k$$

D) Independence of the response error probabilities with respect to time of measurement is also assumed, so that,

(5) $$Q = Q_t \otimes Q_t \otimes \cdots \otimes Q_t,$$

E) Finally, since both conditional independence of response errors over items and independence of time are assumed, there are only L total independent parameters in Q, where;

(6) $$L = \sum_{i=1}^{i=k} (L_i^2 - L_i) = \sum_{i=1}^{i=k} (J_i^2 - J_i) \quad .$$

The construction of $\underline{\Pi}$ is totally dependent on the particular problem being studied. The manner in which the latent class probabilities are expressed is a result of the parameterization chosen, which itself is intended to reflect the structural process and hypothesis being studied.

The Simple Sequence Example: Let us assume that there are four possible stages or categories of cognitive functioning. Each child can be placed in only one stage at any particular time of measurement. Assume further that children are measured at 3 equally spaced points in time. Lastly, the hypothesis of interest is that there is a strict sequence of stages which each child must follow, e.g. Stage I→II→III→IV. One parameterization which we have chosen involves the following parameters:

A) 3 latent initial state parameters $(\alpha_1, \alpha_2, \alpha_3)$ which are the probabilities of the first 3 latent stages. Here the probability of the 4^{th} latent stage at time 1 is denoted by $\alpha_4 = (1 - \alpha_1 - \alpha_2 - \alpha_3) \geq 0$.

B) The remaining parameters are 3 transition probabilities which express the probability of movement among the latent stages: i.e. $P(II|I) = P_1$, $P(III|II) = P_2$ and $P(IV|III) = P_3$.

The transition probability matrix is:

(7)

$$T = \begin{array}{c|cccc} \text{Stage} & \text{I} & \text{II} & \text{III} & \text{IV} \\ \hline \text{I} & 1-P_1 & 0 & 0 & 0 \\ \text{II} & P_1 & 1-P_2 & 0 & 0 \\ \text{III} & 0 & P_2 & 1-P_3 & 0 \\ \text{IV} & 0 & 0 & P_3 & 1 \end{array} \quad \begin{array}{l} \text{Time 2} \\ \\ \\ \\ \end{array} \qquad \text{Time 1}$$

The zeroes in this matrix, T, serve to express the structural hypothesis:

A) that there is no "regression," and

B) that there is no "stage-jumping."

For this model, since there are 4^3 response patterns possible,

(8)

$\underline{P}$ is the 64×1 vector, of the manifest probabilities

Q is a 64×64 matrix of the form

(9) $$Q = Q_1 \otimes Q_1 \otimes Q_1,$$

Q_1 being the 4×4 matrix of misclassification probabilities which is constant over time.

(10) $\underline{\Pi}$ is the 64×1 vector of latent class probabilities, where each latent class is one of the 64 possible patterns

of stage membership. That is, isolating the (j, ℓ) element of T by $(t_{j\ell})$ the latent response pattern triple $< k, j, \ell >$ has the probability

$$P(\underline{\rho} = < k, j, \ell >) = \alpha_k \cdot t_{kj} \cdot t_{j\ell} \quad . \tag{11}$$

The total number of parameters in the simple sequence model for four stages is 18; 12 for Q and 6 for $\underline{\Pi}$.

The estimates of these parameters are found by maximizing the likelihood function defined for the qualitative model by means of the multinomial distribution.

The Quantitative Model: The general form of the models for quantitative data is given by (12).

$$y = \underline{\eta} + \underline{\varepsilon} \tag{12}$$

The structure of the model is on $\underline{\eta}$, the true or latent variables:

$$\underline{\eta} = A\underline{\eta} + \underline{\theta}, \text{ where } \underline{\theta} \text{ is a vector of random variables.} \tag{13}$$

$$\underline{\eta} = (I - A)^{-1}\underline{\theta} \quad . \tag{14}$$

A, $\underline{\eta}$ and $\underline{\theta}$ can be treated as partitioned by time period, each period having multiple variables at each time.

(15) A) $\underline{\theta}_j \sim N(\underline{\mu}_j, \ \Phi_j)$: $J = 1,2,\ldots, n$;

B) $\underline{\eta}_j$, $\underline{\theta}_j$ are $m \times 1$ vectors, i.e. there are m variates at each time.

C) Cov $(\underline{\theta}_j, \underline{\theta}_j') = 0$, for $\underline{j \neq j'}$

D) A_{ij} is an $m \times m$ matrix.

The nature of the matrix A is determined by the structure or hypothesis being investigated. If A is restricted to be lower triangular, then there is no feedback and the model is referred to as a Lag Model. This type of model is most often used for longitudinal data where K, the degree of the Lag, is defined by

(16) $$\underline{\eta}_{t+1} = \sum_{j=1}^{t} (t-k) A \underline{\eta}_j + \underline{\theta}_{t+1},$$

where j denotes the j^{th} time point. The general structure of $\sum_y$, the covariance of the observed variables is given by (17)

(17) $$\sum_y = (I - A)^{-1} \Phi (I - A')^{-1} + \Psi,$$

where Ψ is a diagonal matrix of the error (ϵ) variances, and Φ is the covariance matrix of the random variables $\underline{\theta}$.

The Disengagement Example: The measures of social involvement and psychological involvement will be denoted by y_2 and y_1 respectively, and assume there are three equally spaced observations over time on both variables. The structural hypothesis is that the latent or true amount of social involvement at time $(t + 1)$ is a linear function of the true level of psychological involvement at time (t).

(18)

$y_{1i} = \eta_{1i} + \epsilon_{1i}$, where $i=1,2,3$, and denotes time

$\eta_{1(i+1)} = \alpha_1 \eta_{1i} + \theta_{(i+1)}$, a lag of degree 1.

$\underline{y}_k = A_1 \underline{\eta}_1 + \underline{\epsilon}_1$,

$$A_1 = \begin{bmatrix} 1 & 0 & 0 \\ \alpha_1 & 1 & 0 \\ \alpha_1\alpha_2 & \alpha_2 & 1 \end{bmatrix}$$

(19)
$$y_{2i} = \eta_{2i} + \varepsilon_{2i}$$

$$\eta_{2(i+1)} = \lambda_i \eta_{1i} + \beta_i \eta_{2i} + \theta_{(i+1)}$$

$$\underline{y}_2 = A_{12}\underline{\eta}_1 + A_2\underline{\eta}_2 + \underline{\varepsilon}_2$$

$$A_{12} = \begin{bmatrix} 0 & 0 & 0 \\ \lambda_1 & 0 & 0 \\ \lambda_1\lambda_2 & \lambda_2 & 0 \end{bmatrix}$$

(20)
$$A_2 = \begin{bmatrix} 1 & 0 & 0 \\ \beta_1 & 1 & 0 \\ \beta_1\beta_2 & \beta_2 & 1 \end{bmatrix}$$

(21)
$$\underline{y} = \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix}, \ \underline{\eta} = \begin{bmatrix} \underline{\eta}_1 \\ \underline{\eta}_2 \end{bmatrix}, \ \underline{\theta} = \begin{bmatrix} \underline{\theta}_1 \\ \underline{\theta}_2 \end{bmatrix}$$

$$A = \begin{bmatrix} A_1 & 0 \\ A_{12} & A_2 \end{bmatrix}$$

$$\underline{y} = (I - A)^{-1}\,\underline{\theta} + \underline{\varepsilon}$$

(22)
$$\sum_y = (I - A)^{-1}\,\Phi\,(I - A')^{-1} + \Psi\ ,$$

where $\Phi + \Psi$ are diagonal.

There are a total of $[(6\cdot7)/2] = 21$ statistics in $\sum_y$. There are 6 parameters in $\underline{A}$, $(\alpha_1, \alpha_2, \lambda_1, \lambda_2, \beta_1, \beta_2)$, 6 parameters in Φ $(\sigma^2_{\theta 11},\ldots,\sigma^2_{\theta 23})$, and 6 parameters in Ψ , $(\sigma^2_{\varepsilon 11},\ldots,\sigma^2_{\varepsilon 23})$.